DOCUMENT RESUME

ED 119 724 IR 003 196

| | |
|---|---|
| AUTHOR | Williams, Martha E., Ed. |
| TITLE | Cost Elements and Charge Bases in Information Centers: Proceedings of Panel Discussion. |
| INSTITUTION | Association of Scientific Information Dissemination Centers. |
| PUB DATE | 7 Mar 73 |
| NOTE | 27p.; Paper presented at the Annual Meeting of the Association of Scientific Information Dissemination Centers (Philadelphia, Pennsylvania, March 7, 1973) |
| AVAILABLE FROM | ASIDIC Secretariat, P. O. Box 8105, Athens, Georgia 30601 ($2.50) |
| EDRS PRICE | MF-$0.83 HC-$2.06 Plus Postage |
| DESCRIPTORS | Conference Reports; Costs; *Data Bases; *Fees; *Information Centers; Information Dissemination; Information Services; Questionnaires |

ABSTRACT

At the March 7, 1973, Association of Scientific Dissemination of Information Centers meeting in Philadelphia, Pennsylvania, a panel discussion on cost elements and charge bases in information centers was held. The panel was composed of representatives of seven information centers which process machine-readable bibliographic data bases and provide selective dissemination of information and/or retrospective searches. In order to obtain comparable data and avoid differences in terminology, an advance questionnaire had been completed by the panelists. The questions centered around seven topics: selection of data bases, data base reformatting, system features, data base costs, profile costs, center costs and accounting, and charges and pricing. After each topic was introduced by the moderator the panelists presented their answers. (Author/LS)

**********
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
**********

ED119724

# COST ELEMENTS AND CHARGE BASES IN INFORMATION CENTERS:

# PROCEEDINGS OF PANEL DISCUSSION

ASIDIC Meeting

Philadelphia, Pennsylvania

March 7, 1973

Edited by Martha E. Williams

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Published by the ASIDIC Cooperative Data Management Committee. Copies may be purchased from ASIDIC for $2.50. Checks should be made payable to ASIDIC. Write M.E. Williams, Information Retrieval Research Laboratory, 5-101 CSL, University of Illinois, Urbana, Illinois 61801.

IR 003 196

ASIDIC Secretary
P. B. Schipma
IITRI
10 W. 35th St.
Chicago, Ill. 60616

## PREFACE

At the March 7, 1973 Association of Scientific Dissemination of Information Centers meeting in Philadelphia, Pennsylvania a panel on "Cost Elements and Charge Bases in Information Centers" was held. The panel was arranged and conducted in a manner designed to obtain comparable and compatible information about cost elements or cost factors and the bases upon which prices are determined in seven different information centers. The centers represented and their representatives were:

CAN/SDI (Canadian Selective Dissemination of Information), National Science Library, Ottawa, Canada--Georg Mauerhoff
GIDC (Georgia Information Dissemination Center), University of Georgia, Atlanta, Georgia--Margaret Park
IBM-ITIRC (IBM Technical Information Retrieval Center), Armonk, New York --Herb Nobles
IITRI (IIT Research Institute), Computer Search Center, Chicago, Illinois--Peter Schipma
NC/STRC (North Carolina Science and Technology Research Center), Research Triangle Park, North Carolina--Peter Chenery
NERAC (New England Research Applications Center), University of Connecticut, Storrs, Connecticut--Daniel Wilde
UCLA-CCN (Campus Computing Network), University of California at Los Angeles, Los Angeles, California--Bruce Briggs

The panel discussion was carried out over a three hour time period and was well received by the attendees. It was at the request of many members of the audience that the panel proceedings were put in writing for distribution.

In order to avoid the problem of comparing apples and oranges i.e., measuring different phenomena in different ways, and using different terminology to obtain comparable data, I, as moderator, prepared a set of questions and sent them to each of the panelists in advance of the meeting so that each panelist would be prepared with accurate and ready answers. Answers were to be provided, wherever possible, employing the terms and units of measure used in the questions. This was done to enable the audience to see the differences and similarities of various centers.

For purposes of this panel discussion an information center was considered to be a center that is processing machine-readable bibliographic data bases and providing Selective Dissemination of Information (SDI) and/or retrospective searches from those data bases. Each center represented was processing multiple data bases and, at the time of the panel, had been in operation for at least four years. Thus, each center represented had sufficient experience to provide representative and meaningful data.

Despite the efforts to obtain compatible data there are a number of instances where the differences in center operations, services, data gathering techniques, or accounting practices are responsible for significant differences in the data reported. Where possible I have tried to point out any qualifications in my commentaries preceding the responses to given questions. Questions are also provided as they were asked in order to reproduce as faithfully as possible, the transactions of the panel. Where answers to questions were provided in the form of numerical data, the replies are tabulated to save space.

I wish to thank all of the panelists for their time and effort in preparing responses for the panel discussion and for submitting them to me in writing. The information they have provided will be of interest to many who are concerned with data base services.

Martha E. Williams

Martha E. Williams
ASIDIC Cooperative Data
Management Committee

## SELECTION OF DATA BASES

MODERATOR: WILLIAMS
Some criteria used for, or affecting, the selection of data bases by centers are journal coverage of the data base, currency, multidisciplinariness, comprehensiveness, anticipated marketability, and cost of processing. A user group may wish to be assured that the specific journals they read are covered by a data base if they are going to subscribe to a service from that data base. In some industries currency (time lapse between the appearance of an article in the original journal and the appearance of a reference to the article in the data base) is of prime importance. Multidisciplinariness affords a center the opportunity to provide service to a wide audience from a single data base or from a small number of data bases. Comprehensiveness permits a center to provide in-depth or vertical coverage within a specific field or discipline.

Q1: IN DETERMINING WHICH DATA BASES ARE TO BE PROCESSED AT YOUR CENTER, ARE TOTAL DOLLAR COSTS THE MAJOR CONSIDERATION, OR ARE COST ASSESSMENTS ALSO MADE ON THE BASIS OF UNIT COSTS INCURRED PER RECORD?

CAN/SDI: Mauerhoff
Total cost and unit cost are both considered before a new data base is acquired and processed, but in relation to other variables such as potential market, existing coverage by other data bases, reputation of the producer, etc.

GIDC: Park
The total cost, the unit cost per record, and the size of the potential user community are considered in selecting data bases for the Georgia Center.

IBM: Nobles
Major factors in data base selection are: (1) ability to satisfy the needs of a sizable group of IBM engineers and scientists, (2) currency, (3) completeness of coverage, (4) availability of abstracts on tape, and (5) price.

IITRI: Schipma
Both total cost and unit cost per record are considered but total dollar cost is currently the major consideration.

NC/STRC: Chenery
Total dollar cost is a major consideration.

NERAC: Wilde
If we think a data base can produce enough income to pay for itself, cost is not a consideration. Of course, low cost makes decisions easier.

UCLA: Briggs
Total dollar cost primarily; other factors are considered.

Q2: IN CHOOSING DATA BASES, IS ANY ATTEMPT MADE TO JUDGE JOURNAL COVERAGE IN RELATION TO PROBABILITY OF USAGE, ACCORDING TO BRADFORD DISTRIBUTIONS?

CAN/SDI
Bradford distributions are applicable during acquisition of the first few data bases as the best core coverage is sought, but they become less applicable thereafter since the number of data bases per subject area to choose from is extremely limited, and in most cases is usually only one data base per subject area.

GIDC
No work has been done in this area other than consultations with subject specialists who are familiar with the corresponding printed publications.

IBM
Yes. From data bases such as COMPENDEX we select journals most relevant to IBM. Then periodically review hits per journal to determine changes in selection criteria.

IITRI
No.

NC/STRC
No.

NERAC
Yes.

UCLA
Yes.

Q3: IN TERMS OF OVER-ALL COST EFFICIENCY, DO SYSTEM DESIGN, PROGRAMMING AND OPERATING COSTS FAVOR THE USE OF ONE, OR AT MOST, A FEW LARGE, MULTI - DISCIPLINARY DATA BASES, OR A RELATIVELY LARGE NUMBER OF DISCIPLINE-ORIENTED, SMALLER DATA BASES?

CAN/SDI
CAN/SDI's design, programming and operating costs favor a multi data base environment. This environment assumes that large and popular data bases will help to subsidize small and specialized data bases.

GIDC
There is no simple answer as data bases vary significantly in content (e.g., indexing), hence performance, and choices depend in large part on the type of user community to be supported.

IBM
In ITIRC the system design and operating costs do not favor either large or small data bases. Prime consideration is the number of possible users for a particular data base. Programming costs are of course, determined by the complexity of the conversion programs needed to convert the data base to TEXT-PAC format.

IITRI
I do not think we have enough data to answer this yet. At IITRI we programmed for use of multiple data bases on the assumption that such would be necessary for an excellent service.

NC/STRC
We prefer a large number of discipline-oriented data bases.

NERAC
Yes and no. The fewer the data bases, the easier the system design, programming, and strategy design. However, retrospective searching becomes a burden on large files, e.g. Chemical Abstracts.

UCLA
Our system design and programming effort favor neither. We have not experimented with both to test cost-efficiency, but the ability to add or subtract data base coverage and costs incrementally is an important consideration.

## DATA BASE REFORMATTING

MODERATOR
Data bases can be searched in either the distribution format of the data base supplier or in a search format adopted by a center. This decision is affected by factors such as: the availability of supplier produced search software, the availability of data bases in distribution formats that are compatible with available software packages e.g., TEXT-PAC, the number and types of data bases handled by a center, and the reuse of the data base within the center for multiple purposes, e.g., for retrospective search or creation of specialized subset files in addition to the use of the file for SDI. Reformatting represents a real cost, but presumably it provides a dollar savings in the long run.

Q4: DO YOU REFORMAT INCOMING TAPES FOR SDI?

CAN/SDI
Incoming tapes are reformatted for SDI using a MARC II-like format.
GIDC
Yes.
IBM
We reformat those tapes that are not in TEXT-PAC format.
IITRI
Yes, incoming tapes are reformatted on an issue basis.
NC/STRC
Yes.
NERAC
Yes.
UCLA
At present, but we will not after October 1973.

Q5: DO YOU REFORMAT (AGAIN) FOR RETROSPECTIVE SEARCH?

CAN/SDI
The data bases are not reformatted for those data bases that are used only occasionally for retrospective searches, as the searches are conducted using the SDI programs. The data bases that are used with the on-line interactive system are reformatted.
GIDC
No, we use the reformatted SDI tapes. However, SDI issues are concatenated into volumes for retrospective search.
IBM
No.
IITRI
Yes, on a volume basis. This reformatting is done from the product of the SDI reformat, not from the original tape in distribution format.
NC/STRC
No--same format for SDI and retrosearch.
NERAC
No.
UCLA
No.

Q6: DO YOU GENERATE SUBSET TAPES BY DATE, SUBJECT, OR OTHER CRITERIA?

CAN/SDI
Subset tapes are generated by subject, e.g., pollution and merged to create a new data base. A sort key/compression code is attached to each record to facilitate the removal of duplicate references during the merging.

GIDC
The technical capability is readily available, but there have been no occasions to do so.

IBM
On COMPENDEX and NTIS tapes we generate subset tapes using a variety of selection criteria.

IITRI
Yes, by subject, for customers, at negotiated price and only with approval of the data base lessor.

NC/STRC
No.

NERAC
No.

UCLA
No.

Q7: DO YOU INVERT DATA BASE FILES ON TITLE WORDS, KEYWORDS, AND/OR TERMS IN THE ABSTRACT?

CAN/SDI
No.

GIDC
No. Files are sequenced on document records.

IBM
We do not invert files for batch searching but do for on-line interactive (STAIRS) searching. The total record including abstract words is inverted.

IITRI
We invert CA Condensates on the basis of title words and keywords.

NC/STRC
We invert on keywords.

NERAC
No.

UCLA
No.

Q8: WHAT IS THE COST OF REFORMATTING FOR RETROSEARCH PER VOLUME PER DATA BASE? (SPECIFY)

CAN/SDI
Retrosearch is not currently available on CAN/SDI data bases.

GIDC
Conversion times, hence costs, are data base specific. Total conversion time for FY 73 was 79.4 cpu hours on an IBM 360/65.

IBM
Cost not available as a seperate item.

IITRI
For CA*the cost is approximately $200.00 computer time per volume (150,000 records), plus 2 hours personnel time to submit. The price varies with the size of the data base.

NC/STRC
$50.00 to 100.00 per monthly update.

NERAC
We use our own in-house dedicated machine. Therefore, reformatting costs are hard to break out.

UCLA
For the period July-December, 1972 the average cost per citation (not per volume) is: CA--$.0027 and CAIN--$.0053.

Q9: IS THE COST FOR REFORMATTING FILES FOR RETROSEARCH PASSED ON TO USERS IN YOUR CHARGE FOR SEARCHING OR IS IT CONSIDERED A ONE-TIME DEVELOPMENTAL COST?

CAN/SDI
Retrosearch is not available at present on CAN/SDI data bases.

GIDC
It is included in fees.

IBM
Our total department costs are passed on to users (divisions) according to usage.

IITRI
The cost is passed on to users in the fee for retrosearches.

NC/STRC
It is partly amortized in user fees.

NERAC
We try to recover all costs whenever possible based on full production costs.

UCLA
It is passed on to users.

---

* CA refers to CA Condensates throughout this document.

## SYSTEM FEATURES

MODERATOR
In designing a search system there are various features that may or may not be included depending on a center's assessment of the value of the features versus the trade-offs between the cost of providing those features and the benefit of the features (in terms of user satisfaction and service capability gained).

Q10: DO YOU AGGREGATE SEARCH TERMS, I.E., MAINTAIN IN YOUR PROFILE SEARCH LIST ONLY ONE OCCURRENCE OF TERMS USED BY TWO OR MORE USERS?

CAN/SDI
The profile search list is maintained by profile. Only at the time of a search of a data base are profile terms aggregated and inverted.
GIDC
In the master profile file each term is maintained with the particular profile to which it belongs, regardless of duplication. However, at search time, duplicated terms are identified from the whole profile set and searched only once for each occurrence.
IBM
No.
IITRI
Yes.
NC/STRC
No.
NERAC
No.
UCLA
No.

Q11: WHICH DATA ELEMENTS CAN BE SPECIFIED AS SEARCH TERMS IN YOUR SYSTEM?

CAN/SDI
Data elements that can be specified as search terms are source and personal author; corporate author; CODEN; date; geographic area code; language; type of publication; section; treatment of classification code; cited question; and title, description, index term, or title enrichment elements.
GIDC
Any data element on the data base.
IBM
Any data element on the data base.
IITRI
This depends on data elements on the data base. The software allows search of any element on any data base, but some are restricted for search effectiveness reasons, e.g., journal titles are not searched if CODENs are present.
NC/STRC
Index terms, authors, journal names, languages.
NERAC
In Chemical Abstracts we can search on any data element or character string. In other files--only on subject terms.
UCLA
Any data element present on the data base.

Q12: ARE LEFT AND RIGHT TRUNCATION PERMITTED ON ALL DATA ELEMENTS? IF NOT ON ALL ELEMENTS, ON WHICH ELEMENTS IS LEFT OR RIGHT TRUNCATION PERMITTED?

CAN/SDI

Left, right or no truncation is permitted on only the title, description, index line or title enrichment elements. Right truncation is not used with fixed-length elements such as CODEN, language, type of publication, treatment code, etc. Right truncation is used with geographic area codes and classification codes.

GIDC

Data elements are specified to be either Left-anchored (right truncation only) or Free Text (left and right truncation) at the time the data element search table is constructed. The table can be changed as desired without changing the data base (i.e., it is a function of the search system, not the data base).

IBM

Right truncation only is permitted.

IITRI

Left and right truncation is permitted on essentially all data elements except author name, since it is meaningless.

NC/STRC

No truncation available

NERAC

For example CA--any truncation including adjacent, e.g., poly* *vinyl* *chorid* as a single data element.

UCLA

Yes.

## DATA BASE COSTS

MODERATOR

Centers vary with respect to the data bases they handle. Some provide in-depth coverage from a small number of specialized data bases, some provide broad coverage from multicisciplinary data bases, and others attempt to provide all the data bases required by a defined user community. The value of a data base to a center, the marketability of a data base and the cost of obtaining, maintaining, and processing a data base will vary from center to center depending on the philosophy of operation and sources of support employed by a center.

The purchase, lease, or licensing of data bases is one cost factor in operating an information center. The use-charge, access charge, or royalty is another data base related cost. These data base costs may be a minimal or significant portion of an information center operation depending on the data bases used, changes made to the data base, and method of processing.

Q13: WHAT ANALYSIS IS MADE OF SUCH FACTORS AS COMPREHENSIVENESS OF COVERAGE AND CURRENCY OF THE DATA IN DETERMINING COST-BENEFIT VALUES OF EACH DATA BASE CONSIDERED?

CAN/SDI

Comprehensiveness of coverage and currency of the data are very important but usually fall to the wayside in most instances because a center does not enjoy much of a choice of data bases when contemplating the addition of, say a file for engineering. Usually the choices are limited to one, and either you acquire it or not.

GIDC

None.

IBM

Analysis of data is made by specialist personnel and management relative to comprehensiveness and currency of data. Obviously we are concerned primarily with appropriate engineering and data processing information and applications needed by IBM personnel. We restrict input of appropriate data to the current two year period.

IITRI

Our approach is to decide on a data base primarily in terms of its content and our assessment of its responsiveness to the user clientele.

NC/STRC

The final determination in selecting data bases to be searched is made by clients. (we reccommend, they decide)

NERAC

Recommendations are made by the marketing staff.

UCLA

Both items are extremely important factors in the consideration. Information on these factors is obtained both from the supplier and, if possible, from other users.

Q14: WHICH DATA BASES DO YOU PROCESS FOR SDI AND WHAT PERCENTAGE OF YOUR TOTAL OPERATING COSTS ARE SPENT ON DATA BASE SUBSCRIPTIONS; WHAT PERCENTAGE IS SPENT ON ROYALTIES (IF APPLICABLE)?

CAN/SDI
The SDI data bases processed for the entire CAN/SDI Consortium are: BA, BIORI, CA Even, CA Odd, CAIN, COMPENDEX, CT, ERIC, GEO-REF, GRA, INSPEC, MARC II, MEDLARS, SCI, and SSCI. Subscriptions amount to approximately 9% of total operating costs; royalties are almost identical.

GIDC
The SDI data bases are: BA, BIORI, CA Even, CA Odd, CAIN, CIJE, COMPENDEX, GEO-REF, GRA, NSA, PA, RIE, and SPIN. Approximately 7% of the center's costs are for data base subscriptions plus royalties.

IBM
IBM Technical Reports; IBM Invention Disclosures and Patents; NON-IBM Technical Reports (including NTIS); COMPENDEX, plus data processing journals; CAS's POST; and SPIN. Subscription costs are approximately 3% of the total budget. Royalties are negligible.

IITRI
Data bases used for SDI are: CA Even, CA Odd, COMPENDEX, and FSTA. Lease costs are approximately 7% of operating costs, and royalties are approximately 5%.

NC/STRC
ERIC, FSTA, GRA, ITT, NASA, WTA. Data base costs are 4%.

NERAC
We provide SDI from: CA, ERIC, GRA, METADEX, NASA, and WAA. We spend about 1.3% of our total budget on data base subscriptions; about .25% on royalties.

UCLA
BA, BIORI, CA Even, CA Odd, CAIN, COMPENDEX, and ERIC. Subscriptions are 5.1% of total operating costs and royalties are 0.5%.

Q15: AT WHAT LEVEL OF UTILIZATION IS A DATA BASE CONSIDERED TO BE CARRYING ITS OWN WEIGHT IN THE TOTAL CONTRIBUTION TO USE OR REVENUE? ARE LITTLE-USED DATA BASES CANDIDATES FOR TERMINATION OR ARE THEY CARRIED FOR THE DURATION, IMPACTING THE COST SIDE OF THE LEDGER, BUT CONTRIBUTING NEGLIGIBLY TO REVENUE OR USAGE?

CAN/SDI
A rule of thumb has been used which is "100 profiles after one year and a data base is carrying its own weight." Once a decision has been made to acquire a data base, the data base will be processed for the duration barring "acts of God."

GIDC
The selection of data bases is examined annually and decisions made to renew or discontinue. The decisions are based on the user community as a whole and in conjunction with costs.

IBM
A data base is considered cost recoverable at 200 profiles per month plus approximately 1000 searches per year.

IITRI
The latter is true initially--a given data base is carried by other data bases until enough subscriptions are obtained to meet data base cost (royalties, computer time, and out-of-pocket costs). If it then looks as though growth will continue until full costs are met, it is carried another year after which time it is expected to bring in sufficient revenue to cover the related costs.

NC/STRC
We have dropped two data bases for which usage did not justify the fixed cost.

NERAC
A data base should at least cover its own purchase cost.

UCLA
We determine "cost" of data base as follows: c=subscription plus royalties plus operating cost and computer search cost. Then divide c by number of users (profiles) to determine "cost" per profile. If not in line with retro schedule and/or "cost" of other data bases, it is a condidate for termination.

Q16: WHAT PROCEDURES OR TECHNIQUES ARE USED TO EVALUATE THE POTENTIAL MARKET ACCEPTABILITY (IN A USAGE AND/OR REVENUE SENSE) OF ALTERNATIVE DATA BASES WITH THE FULL KNOWLEDGE THAT WIDE USER ACCEPTABILITY IS THE SURE-FIRE WAY TO REDUCE TOTAL UNIT COSTS?

CAN/SDI
Nationally distributed questionnaires prior to the implementation of the service provide information on candidate data bases, as well as possible implementation orders for these data bases. Seminars for intermediaries, feedback from users, requests for extended coverage, etc. all figure into market acceptability. However, although individuals and groups request additional coverage and there are known gaps in the coverage there is no guarantee that they will subscribe when the data bases are actually offered.

GIDC
N/A

IBM
Since all service is rendered internally to IBM employees, the potential acceptability is established by surveys of existing system users, other laboratory and headquarters personnel, and technical library personnel.

IITRI
We survey our current users and potential users known to us via prior contacts or inquiries.

NC/STRC
Market surveys (mail, personal interviews) of present and prospective clients are conducted.

NERAC
Recommendations are made by the marketing staff.

UCLA
We employ feedback from users and reference librarians, combined with staff experience.

## PROFILE COSTS

MODERATOR
There are many ways of looking at the costs associated with the running of profiles against data bases in order to provide SDI services. For example, the average profile size varies both from center to center, and from data base to data base within the center. Record size in the different data bases varies, and there may be additional variance introduced by a given center's adding or deleting material in the records of a particular data base. Profile writing, maintenance and updating can take varying amounts of time depending on the individual data base's data element content, vocabulary, and stability, as well as on the growth rate of the total profile set and the search features available in the software system employed. It is probably easier for one profile writer to maintain 300 profiles on one data base than to maintain 100 each on three different data bases because of the effort involved in becoming familiar with each data base's content and conventions, and of keeping up with data base changes that must be implemented in the profile and communicated to the user. The major burden of profile development and adjustment is put on the user by some centers, while other centers provide extensive user assistance and screening of output.

Q17: WHAT IS THE COST PER RETRIEVAL HIT IN YOUR SYSTEM? PLEASE INDICATE THE VARIANCE FROM DATA BASE TO DATA BASE AND ISSUE TO ISSUE.

CAN/SDI
Machine costs per retrieved hit (revised as per new computer charging scheme) are as follows: SCI--3.6¢; CA--2.4¢; COMPENDEX--2.4¢; INSPEC--3.4¢.

GIDC
Costs are not calculated.

IBM
For SDI the cost is 18¢ per hit. For retrospective search the cost averages $18.00 per search per data base and the average search generates 50 hits.

IITRI
The cost per retrieved hit is approximately 10¢ but standard deviation is very high. For any given data base. variance is high because of profile mix. Between data bases, it is fairly reasonable.

NC/STRC
Average computer cost varies from 3¢ to 5¢ per retrieved hit.

UCLA
I do not have data on cost per retrieved hit, however cost per hit is as follows: CA Odd--5.5¢; CA Even--8.4¢; CAIN--6.3¢.

Q18: WHAT IS YOUR COST, IN TERMS OF MACHINE TIME, FOR MATCHING ONE PROFILE AGAINST 5,000 CITATIONS? INCLUDE THE PROPORTIONATE COST OF FORMAT CONVERSION (EXCLUDE DATA BASE LEASE AND ROYALITIES). NORMALIZE FIQURES FOR A PROFILE OF 10, 20, 30 TERMS. PLEASE INDICATE WHICH DATA ELEMENTS REMAIN ON THE TAPE(S) YOU SEARCH AND INDICATE THE AVERAGE NUMBER OF CHARACTERS PER CITATION IN THE SEARCH FILE.

CAN/SDI

The cost per profile (60 terms) per tape issue of approximately 5,000 citations is $1.49, $1.57, and $1.88 respectively for SCI, CA, COMPENDEX and INSPEC. This includes conversion of the tape retaining all elements, search of converted tape, sort and print. (CAN/SDI uses a 360/65)

GIDC

Search costs in terms of machine time have been calculated for 1 profile (26.6 terms) versus 5,000 citations. The average cpu time for CA (Odd and Even combined) is .046 cpu minutes. These costs do not include format conversion. (GIDC uses a 360/65).

IBM

One memory load (up to 100 profiles) searches approximately 5,000 citations (over 100 words per citation--title, author, corporate author, data accession number, source code, abstracts, categories) in approximately 6 minutes. Average 40 terms per profile. (IBM uses a 360/65 and a 370/155)

IITRI

Assuming 100 profiles are batched for searching, the average cpu time on a 360/65 is: for a profile of 10 terms the time is approximately 6 seconds; for a profile of 20 terms the time is approximately 12 seconds; and for a profile of 30 terms the time is approximately 18 seconds. The data elements on the search tape are title; author(s); company affiliation; keywords; CODEN: bibliographic citation; and journal title. The mean citation length is 282 characters.

NC/STRC

Normal data elements are key terms and accession numbers with an average of 18 terms per citation. (NC/STRC uses a 370/165).

NERAC

For a profile of 10 terms the time is approximately .6 minutes; for a profile of 20 terms the time is approximately .7 minutes; for a profile of 30 terms the time is approximately .8 minutes. We run on an IBM 1130.

UCLA

For a profile of 10 terms the time is approximately .112 for CA and .094 for CAIN; for a profile of 20 terms the time is approximately .223 for CA and .188 for CAIN; for a profile of 30 terms the time is approximately .335 for CA and .282 for CAIN. (figures for Odd and Even issues of CA have been averaged). (UCLA uses a 360/91).

Q19: HOW MANY 20 TERM PROFILES CAN BE HANDLED IN ONE SDI RUN? WOULD THE COST/PROFILE VERSUS A 5,000 CITATION TAPE DECREASE SIGNIFICANTLY AS THE NUMBER OF PROFILES INCREASES?

CAN/SDI
Search phase is system bound and will use however much core is available to it, i.e. 35-307K. Using most popular data base which is ISI Source & Citation, 9 disk passes were required to process data base of 5625 full citations. Each pass ranged in size from 70 to 110 inverted profiles.

GIDC
The number of profiles which can be handled in one SDI run is a function of the amount of core allocated and is therefore a user parameter. The cost/profile decreases rapidly with the addition of more profiles for SDI runs of up to about 10 profiles. After 10 profiles (average), the cost/profile levels off and is essentially constant.

IBM
Cost would decrease significantly when profiles increase form 1 to 100. Then the second memory load costs the same as the first with the cost/profile decreasing from 101 to 200. We currently run 30 memory loads/data base.

IITRI
This depends on amount of core available. Using an IBM 360/65 with 300K core, approximately 200-300 profiles can be accomodated. Cost does not decrease significantly after 100 profiles per batch.

NC/STRC
100 profiles per run. Cost does not decrease as number of profiles is increased.

NERAC
Approximately 20 with our small machine. Decrease in cost will result from more profiles; but amount depends on total job mix.

UCLA
Currently, over 500; restricted only by available core (can get up to 1.2KK bytes).

Q20: HOW MANY SDI PROFILES FOR ONE DATA BASE CAN ONE PROFILE WRITER HANDLE?

CAN/SDI
Using geographically-remote intermediaries, who prepare initial profile searches, CAN/SDI search editors at National Science Library were looking after 1162 SDI profiles or 387 per editor. These same editors were also engaged in retrosearching on non CAN/SDI systems, user education and training, etc. CAN/SDI was running 10 data bases. Full time intermediaries who are responsible for all their user contacts, interviewing, training, initial profile creation, updating, monitoring, etc. could probably handle more than 100 profiles.

GIDC
The Georgia Center currently processes over 9,700 SDI profiles on 12 data bases which are handled by 3.5 full time equivalents of information scientists (profile writers). The average profile is run against 2.7 data bases, thus one profile writer handles approximately 1000 unique profiles.

IBM
Each profile writer (Information Retrieval Specialist) maintains roughly 500 profiles.

IITRI
Approximately 300 if all profiles are written for one data base; 200-250 if for two or three data bases.

NC/STRC
35-50.

NERAC
Depends on his technical background. A metallurgist should be able to write many more searches on METADEX than say an educator. On the other hand, the educator should do better on ERIC. I would like to see our profilers generate about 4 searches a day, including talking to the clients and reviewing output.

UCLA
NA. Our profile preparation is handled by reference librarians in branch libraries.

Q21: HOW MANY DATA BASES CAN ONE PROFILE WRITER KEEP UP WITH?

CAN/SDI
Using geographically-remote intermediaries, who prepare initial profile searches, CAN/SDI search editors at National Science Library were looking after 1162 SDI profiles or 387 per editor. These same editors were also engaged in retrosearching on non CAN/SDI systems, user education and training, etc. CAN/SDI was running 10 data bases. Full time intermediaries who are responsible for all their user contacts, interviewing, training, initial profile creation, updating, monitoring, etc. could probably handle more than 100 profiles.

GIDC
All profile writers prepare profiles for the 20 data bases used in the center although data bases tend to be grouped by subject specialty area, hence profile writer specialty. Thus, each profile writer usually knows about 5 or 6 data bases very well and the others sufficiently well to handle occasional questions.

IBM
Each specialist manages one data base. Data base assignments are rotated each two years.

IITRI
Four or five and the effort's reduced if one or two are subsets for another, e.g., POST and CBAC are subsets of CA.

NC/STRC
12-15

NERAC
This depends on the profile writer's technical background. A metallurgist should be able to write many more searches for the METADEX data base than say an educator. On the other hand, the educator should do better on ERIC. I would like to see our profilers generate about 4 searches a day, including talking to the clients and reviewing output.

UCLA
At UCLA profile preparation is handled by reference librarians in branch libraries and not by the center per se.

Q22: WHAT IS THE AVERAGE NUMBER OF TERMS PER PROFILE PER DATA BASE IN YOUR SYSTEM NOW AND PREVIOUSLY? IF THE NUMBER HAS NOT REMAINED CONSTANT, TO WHAT DO YOU ATTRIBUTE THE VARIANCE? HAS IT STABILIZED TO THE POINT WHERE YOU CAN MAKE ACCURATE PREDICTIONS REGARDING THIS ELEMENT OF DATA AS A COST FACTOR?

CAN/SDI

Average number of terms per profile per data base is 67.4. A profile subscription entitles the user to 60 terms.

GIDC

Overall, the average number of terms per profile is 26.6 with a standard deviation of 31.4. The minimum number of terms per profile is 1 and the maximum is 255.

IBM

Average 40 terms per profile. Since search logic as well as number of terms effects search time, this factor is difficult to evaluate as an individual cost element.

IITRI

An average of 29 terms per profile has been a fairly stable number for three years. The range is 1 to 200+.

NC/STRC

An average of 40 terms per search.

NERAC

Our strategy designers average approximately 20 terms per search. A beginning strategy designer usually uses more until he learns the data base. He usually tries tight strategies and produces few if any hits. On chemical searches, we average 12-15. Probably less because of the truncation, but truncation produces longer search times. Here, we have traded off strategy designer time for computer time.

UCLA

The average for CA and CAIN is approximately 30. The variance has been slight.

## CENTER COSTS AND ACCOUNTING

MODERATOR
The costs associated with center operations vary from center to center. Centers must find ways of determining cost and know what units are to be measured for cost assignment purposes. Cost recording practices should provide a basis for the management information a center needs, and cost analysis is necessary if the cost accounting system is to have an impact on the total system. In some cases the manner of recording data and accounting for personnel is governed by the type of accounting system that is imposed on a center from the parent organization. Overhead rates and salary related costs (benefits) differ. Computer charges differ and are associated with different cost items; e.g., cpu time, cards read or punched, tapes mounted and read, wall clock time, etc.

Q23: DO YOU HAVE YOUR OWN ACCOUNTING SYSTEM OR ARE YOU REQUIRED TO USE THE SYSTEM DEVELOPED BY YOUR ORGANIZATION? (PLEASE EXPLAIN)

CAN/SDI
CAN/SDI utilizes its own accounting system to keep track of profiles, terms, overterms, cited questions, etc. and generates its own "work-performed" summaries. These summaries are then forwarded to the National Research Council's Financial Services which sends invoices out to the users.

GIDC
The accounting procedures include data collection as part of the automated billing component of the text search system with direct input to the Computer Center's accounting system and subsequent automatic input to the University's financial systems.

IBM
We have our own system.

IITRI
We maintain our own computer cost accounting system, but that cost plus personnel and materials charges are all handled by IITRI's overall accounting routine.

NC/STRC
We have our own system.

NERAC
We have our own system.

UCLA
We have our own system.

Q24: PLEASE INDICATE THE RELATIVE COST FIGURES FOR THE FOLLOWING IN RELATION TO THE TOTAL COST OF OPERATING YOUR SYSTEM.

TABLE OF RELATIVE COSTS FOR CENTER OPERATIONS

| COSTS (%) | CENTERS | | | | | | |
|---|---|---|---|---|---|---|---|
| | CAN/SDI CY72* | GIDC FY73* | IBM CY72* | IITRI[a] CY72* | NC/STRC CY72* | NERAC CY72* | UCLA CY72* |
| Data base lease/license/ royalty and address fees | 18.0 | 7.0 | 3.0 | 10.0 | 4.0 | 2.0 | 5.5 |
| Purchased expendable materials (cards, tapes, etc.) | 2.0 | 11.9 | 2.0 | 5.0 | 3.0 | 0.8 | 1.3 |
| Postage; shipping cost | 1.5 | 0.1 | 1.0 | 5.0 | 1.0 | 0.6 | 0.7 |
| Capital purchase--equipment, etc. (specify) | 3.0 | -0- | -0- | -0- | -0- | -0- | 3.0[c] |
| Computer time: | | | | | | | |
| SDI | NA | NA | 20.0 | 15.0 | 0.2 | 0.7 | 9.4 |
| Retrosearch | NA | NA | 10.0 | 3.0 | 4.0 | 1.4 | 1.4 |
| Test programs, R & D | NA ** | NA ** | 2.0 | 1.0 | 2.8 | 1.4 | 8.1 |
| Tape copy, conversion, and other maintenance | NA | NA | 1.0 | 1.0 | 0.5 | 7.7 | 0.5 |
| See Note b | NA | NA | -0- | -0- | -0- | 2.8 | -0- |
| Total Computer Time | 18.0 | 37.0 | 33.0 | 20.0 | 7.5 | 14.0 | 19.4 |
| Personnel: | | | | | | | |
| Management | 8.0 | 16.6 | 8.0 | 5.0 | 8.0 | 3.9 | 9.1 |
| Marketing, education, training, materials | NA | 3.7 | 1.0 | 10.0 | 19.5 | 14.4 | 3.8 |
| System design, programming and operations | NA [d] ** | 2.9 | 6.0 | 3.0 | 12.0 | 2.8 | 40.3[h] |
| Profile handling | NA | 16.3 | 10.0 | 25.0 | 30.0 | 12.4 | 8.2[d] |
| Clerical, keying, distrib | NA | 3.1 | 3.0 | 15.0 | 14.5 | 5.1 | 2.7 |
| Maintain tape library | NA | 0.4 | 5.0 | 2.0 | 0.5 | 0.4 | 0.4 |
| Total Personnel | 29.0 | 43.0[g] | 33.0 | 60.0 | 84.5 | 39.0 | 64.5[e] |
| Other | 28.5 | 1.0 | 28.0[f] | -0- | -0- | 43.6 | 5.6 |
| TOTAL | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

[a] Based on CA Condensates
[b] Use of wholly owned computer for non-search related work
[c] Including equipment rental
[d] Most done by reference librarians and profile writers outside of the center
[e] Additional personnel allocation, 2.69 for operations
[f] Microprocessing, hard-copy production and distribution, library bulletins, etc.
[g] Based on average rates for personnel in listed job categories
[h] During CY72, UCLA was still in the design phase

* FY = fiscal year; CY = year during which the majority of the calendar months occurred
** -0- means answer given was zero; NA means not available

Q25: HOW IS COMPUTER TIME CHARGED IN YOUR INSTITUTION?

CAN/SDI
Computer time costs $4.50 per cpu minute plus $0.75 per 1000 lines of print.
GIDC
There is a billing algorithm used which takes into account core, channel usage, I/O units, cpu time, etc. All rates are subject to Federal audit.
IBM
Departmental charges according to usage algorithm.
IITRI
By cpu second and I/O operations and materials used.
NC/STRC
Actual usage: cpu, EXCP, I/O
NERAC
We have a fixed cost computer dedicated to our center.
UCLA
Rate is $ .11/MUS where MUS = (T + .02I) (1 + .0036 (MIN(12,250) ) +.0004R)
T = cpu time, I = I/O count, R = region size in kilobytes

Q26: ARE YOU ABLE TO IDENTIFY CUP TIME VERSUS I/O VERSUS WALL CLOCK TIME?

CAN/SDI
We can identify cpu time and wall clock time.
GIDC
Yes.
IBM
Times are identified for each job on the JCL class A printout.
IITRI
Yes.
NC/STRC
Yes.
NERAC
We cannot separate output time from I/O time except by substracting off time to do a straight tape read.
UCLA
Yes.

Q27: DO YOU PAY FOR YOUR COMPUTER TIME FROM THE PROFILE FEES?

CAN/SDI
No.
GIDC
Indirectly.
IBM
All costs are compiled and recovered through all service charges.
IITRI
Yes.
NC/STRC
Yes.
NERAC
Yes.
UCLA
Not entirely.

Q28: IS PRINTING DONE OFF-LINE OR ON-LINE?

CAN/SDI
On-line.

GIDC
Both, but the bulk is done off-line on IBM 1401s and a CDC 6400.

IBM
Off-line for batch processing with optional off-line/on-line for STAIRS terminal system.

IITRI
On-line, but through HASP, so it is effectively off-line.

NC/STRC
On-line.

NERAC
On-line.

UCLA
On-line.

Q29: PLEASE INDICATE THE RELATIVE PERCENTAGES (VERSUS TOTAL COMPUTER COSTS FOR THE SDI ACTIVITY) OF COMPUTER COSTS REQUIRED FOR:

| PROGRAM FUNCTION | CENTER AVERAGES | | | | | | |
|---|---|---|---|---|---|---|---|
| | CAN/SDI* | GIDC | IEM | IITRI | NC/STRC | NERAC | UCLA |
| Format conversion of data base | 6 | NOT AVAILABLE | 15 | 40 | 45 | 70 | 20 |
| Profile handling | 13 | | 9 | 14 | 5 | -- | 6 |
| Search | 59 | | 71 | 40 | 10 | 20 | 44 |
| Output preparation including printing | 20 | | 1 | 5 | 35 | 10 | 30 |
| Accounting and statistics and automatic report generation | 2 | | 4 | 1 | 5 | -- | |
| TOTAL | 100 | -0- | 100 | 100 | 100 | 100 | 100 |

"RELATIVE PERCENTAGES OF COMPUTER COSTS PER PROGRAM FUNCTION FOR EACH CENTER"

* Percentages vary considerably among data bases.

## CHARGES AND PRICING

MODERATOR
The start-up costs (design, programming, and debugging) for most centers have been covered by grants from the federal or state sources. Some have required that a center become self supporting within a given time period and others have not. Presumably, start-up costs are incurred once and could be allocated to the first "X" years of operation. Center services are marketed in different ways depending on the type of user clientele a center wishes to service. This affects the total operating cost which, in some cases, is reflected in the charges levied for services.

Q30: WHAT IS THE COST IN TERMS OF MAN HOURS OF EFFORT REQUIRED TO SELL A PROFILE WITHIN YOUR ORGANIZATION AND OUTSIDE YOUR ORGANIZATION? PLEASE INDICATE THE LEVEL OF PERSONNEL USED FOR THIS TYPE TASK. WHAT METHOD IS MOST EFFECTIVE?

CAN/SDI
Cost in terms of man hours of effort required to sell a profile is not available since geographically-remote intermediaries perform much of the "selling". Intermediaries are librarians and information officers. The most effective method is by word-of-mouth advertising although on-site seminars and demonstrations have proved useful.

GIDC
The Georgia Center has no marketing program per se. Promotion is by seminar, personal contacts, etc. within the University System of Georgia.

IBM
Management and specialists (professional level) share the responsibility for increasing the user population. Approximately 5% of management time and 10% of specialist time is required for this effort. The best approach, to date, has been educational seminars at laboratory locations and scientific centers explaining ITIRC's services and signing up users on the spot.

IITRI
Within IITRI 1-10 hours and outside, 5-20 hours. Promotion or marketing is done by high level technically competent personnel.

NC/STRC
Information not available.

NERAC
We only sell searches outside our organization. The time required to sell a search varies from 15 minutes to several days depending upon the client. In general, we do not sell single searches because the cost of marketing is too high. How can an organization afford to send out a salesman to generate a $200.00 search even if he sold one on every call?

UCLA
NA.

Q31: ARE START-UP COSTS ACCOUNTED FOR IN ESTABLISHING USER CHARGES?

CAN/SDI
Start-up costs have not been considered in establishing user charges.

GIDC
Start-up costs which were not federally funded are amortized into the costs and user charges.
IBM
ITIRC start-up costs (system design and implementation) was covered by Corporate Headquarters on a non-recovery basis.
IITRI
No.
NC/STRC
No.
NERAC
Yes.
UCLA
No.

Q32: WHAT ARE THE BASES ON WHICH YOU DETERMINE CHARGES?

a) NUMBER OF SEARCH TERMS
b) CHARACTERISTICS OF SEARCH TERMS (HIGH FREQUENCY VERSUS LOW FREQUENCY: SUBJECT VERSUS AUTHOR, ETC.)
c) NUMBER OF HITS GENERATED
d) NUMBER RECORDS PRINTED AND SENT TO USER
e) FLAT FEE PER DATA BASE
f) FLAT FEE FOR ANY DATA BASE
g) OTHER (EXPLAIN)

CAN/SDI
Charges are determined by: (a) number of search terms; (b) frequency of output; (c) mailing charges including envelopes; (d) computer charges; (e) royalties and data base charges.
GIDC
The charges are determined on the basis of the center's operational costs and are generally based on a fixed fee per data base per volume or issue searched. Additional charges are assissed for optional services, such as printing of abstracts, special print stock, etc.
IBM
Charges are based on hits.
IITRI
a) Yes. A maximum number of search terms is allowed within a profile subscription. An added charge is levied for additional terms.
b) No.
c) No. A maximum number of prints are allowed with a profile subscription. There is a fee for prints in excess of this number.
d) Yes.
e) A flat fee or base fee is charged for a maximum number of terms and prints. Added charges are made for extra term or print units.
f) No.
g) --
NC/STRC
a) This applies to wholesale clients.
b) This applies to wholesale clients.
c) --
d) This applies to wholesale clients.
e) This applies to retail clients.

NERAC
Flat fee per data base with discount for search of multiple data bases on same subject.
UCLA
SDI: per profile per data base; retro: per profile per volume of data base.

Q33: ARE OVERHEAD COSTS CONSIDERED IN ESTABLISHING YOUR PRICING SCHEDULE?

CAN/SDI
No.
GIDC
Yes.
IBM
Yes.
IITRI
Yes.
NC/STRC
No.
NERAC
Yes.
UCLA
Yes.

Q34: PRESUMABLY ALL OF THE CENTERS ARE WORKING TOWARD, OR HAVE ACHIEVED, A BREAK EVEN POINT FOR COVERING THEIR OPERATING COSTS AND SOME MAY EVEN LOOK TO THE POSSIBILITY OF PROFIT. IN ORDER TO ACHIEVE THESE GOALS, WHICH OF THE FOLLOWING THREE POSSIBLE TACKS ARE YOU TAKING: (1) INCREASING THE NUMBER OF SUBSCRIPTIONS: (2) INCREASING THE CHARGE FOR SUBSCRIPTIONS: (3) MAKING UP THE DEFECIT BY PUBLIC SUBSIDY?

CAN/SDI
While library subsidy will continue to be involved, an increase in the number of subscriptions is sought to cover more and more of the operating costs.
GIDC
The basic operations of the Center are supported as a University System service, much as are the library facilities. Search services are offered to non-University System users on a cost-recovery basis. The question is not appropriate to our operation.
IBM
Since we seek a "break even" operation the number of users as reflected in hits and searches (plus hard copy, microfiche, etc.) directly effects unit charges. More usage means lower unit price (SDI, searches, etc.).
IITRI
No. 1, hopefully; no. 2, if necessary.
NC/STRC
All three but no. 3 for initial period only.
NERAC
All three.
UCLA
No. 1.